全基因组关联研究中极端不平衡数据的统计分析方法

谢宁[1#]，毕文健[2#]，张中文[1]，邵方[1]，魏永越[3]，赵杨[1,2,4]，张汝阳[1,4,5*]，陈峰[1,4*]

[1]南京医科大学公共卫生学院生物统计学系，南京 211166

[2]北京大学基础医学院医学遗传学系，北京 100191

[3]北京大学公众健康与重大疫情防控战略研究中心，北京 100191

[4]南京医科大学环境与人类健康国际联合研究中心，南京 211166

[5]南京医科大学常州医学中心，常州 213164

【摘要】极端不平衡数据定义为自变量或因变量指标的取值呈现严重比例失衡的数据，例如病例-对照极度不平衡、疾病发病率极低、生存数据大量删失以及遗传位点为低频或罕见变异等。在此情境下，logistic 回归模型、Cox 比例风险模型等参数假设检验的经典统计量偏离正态分布，难以控制一类错误。近年来，随着超大型人群队列全基因组关联研究资源的日益共享与深度挖掘，高效准确处理独立或非独立样本极端不平衡数据的统计需求日益突出。为此，本文系统地进行了方法学概述。首先，综述常见经典统计量理论推导的原理；其次，阐述极端不平衡数据对统计量分布的影响；然后，介绍遗传统计学中常用的两种统计量校正方法：Firth 校正和鞍点近似方法；最后，简介极端不平衡基因组学数据常用软件。本文为极端不平衡数据的统计分析提供理论参考和应用推荐。

【关键词】全基因组关联研究；极端不平衡数据；Firth 校正；鞍点近似；罕见变异。

近年来，随着大型生物样本库和全基因组测序数据的日益释放与共享，遗传数据深度挖掘的机遇与挑战并存。一方面，更加充足的样本和测序深度有助于识别新的疾病相关标记物；另一方面，采用常规统计方法分析极端不平衡数据(extremely unbalanced data)可能会导致严重的一类错误膨胀[1-3]。全基因组关联研究(genome-wide association study, GWAS) 涉及百万、千万个位点，如果无法严格控制一类错误，那么假阳性位点的绝对数量会很大，将对后续研究造成严重干扰和资源浪费。GWAS 极端不平衡数据主要体现在：(1) 自变量是低频甚至罕见变异，例如次等位基因频率(minor allele frequency, MAF) < 0.01 或英国生物标本库(UK Biobank, UKB)级别数据中次等位基因计数(minor allele count, MAC) < 400 的位点；(2) 因变量是二分类数据，但事件的比例较低。例如，在病例-对照研究中，病例与对照比例极度不平衡，超过 1:20；在队列研究中，疾病的发病率较低，低于 5%；(3) 因变量是生

存数据，但删失比例严重，超过 90% [3, 4]。因此，本文中极端不平衡数据定义为自变量和/或因变量指标的取值呈现严重比例失衡的数据。logistic 回归模型、Cox 比例风险模型等参数假设检验的经典统计量的应用前提是：大样本条件下，统计量收敛于正态分布或卡方分布。然而，对于极端不平衡数据，即便在大样本条件下，经典统计量仍然无法有效收敛于对应分布，难以控制一类错误[5]。

本文内容结构如下：第一节回顾常见统计模型参数假设检验的理论基础；第二节阐述经典统计方法无法用于极端不平衡数据统计分析的理由；第三节着重介绍 Firth 校正和鞍点近似(saddlepoint approximation, SPA)方法，阐述两种方法处理极端不平衡数据的原理和效果。第四节简介基因组学极端不平衡数据常用软件。

一、经典模型参数假设检验的理论方法回顾

1.1 经典的大样本检验

首先，我们回顾三种渐近等效的大样本假设检验方法，即 Wald 检验、似然比检验和得分检验。这三种方法以极大似然估计和中心极限定理为依据，构造大样本下服从正态分布或卡方分布的检验统计量，在参数模型的假设检验中起核心作用。

(1) Wald 检验

当样本量较大时，总体参数γ的极大似然估计$\hat{\gamma}$在适当的正则条件下会依分布收敛于正态分布[6]，即：

$$\frac{\hat{\gamma}-\gamma}{\mathrm{se}\{\hat{\gamma}\}} \xrightarrow{d} N(0,1) \tag{1}$$

式(1)中，$\mathrm{se}\{\hat{\gamma}\}$是极大似然估计$\hat{\gamma}$的标准误。对于极大似然估计，相应的$\mathrm{se}\{\hat{\gamma}\}$具有性质：随着样本量增加，收敛于$1/\sqrt{I_n(\gamma)}$。此处的$I_n(\gamma)$被称为期望信息量或 Fisher 信息量[7]。设$X_1,\ldots,X_n$来自于总体分布$f(x|\gamma)$的随机样本，则$I_n(\gamma)$的具体公式如下：

$$\begin{aligned} I_n(\gamma) &= E\left(\frac{\partial}{\partial\gamma}\log L(\gamma|\boldsymbol{x})\right)^2 \\ &= -E\left(\frac{\partial^2}{\partial\gamma^2}\log L(\gamma|\boldsymbol{x})\right) \qquad (\text{对于指数族恒成立}) \\ &= -E\left(\sum_{i=1}^{n}\frac{\partial^2}{\partial\gamma^2}\log f(x_i|\gamma)\right) \\ &= -nE\left(\frac{\partial^2}{\partial\gamma^2}\log f(x_i|\gamma)\right) \end{aligned} \tag{2}$$

式(2)中的$L(\gamma|\boldsymbol{x})=\prod f(x_i|\gamma)$为似然函数，log 为以 e 为底的对数。由于真实的总体参

数γ往往是未知的，因此式(2)中的γ常用极大似然估计$\hat{\gamma}$进行替代，最终得到$\mathrm{se}\{\hat{\gamma}\} \approx 1/\sqrt{I_n(\hat{\gamma})}$。

由此若需要检验假设$H_0: \gamma = \gamma_0$，则可构造式(3)中的 Wald 检验统计量，当H_0成立时，该统计量在大样本下服从标准正态分布：

$$Z_{\text{Wald}} = \frac{\hat{\gamma} - \gamma_0}{\mathrm{se}\{\hat{\gamma}\}} \tag{3}$$

以显著性水平α进行双侧检验时，若$|Z_{\text{Wald}}| > Z_{1-\alpha/2}$则拒绝$H_0$接受备择假设。$Z_{1-\alpha/2}$表示标准正态分布的(1-$\alpha/2$)×100%分位数。同样，通过$\hat{\gamma} \pm 1.96se\{\hat{\gamma}\}$求得参数$\gamma$的 95%置信区间。将 Wald 统计量取平方可得到卡方统计量的形式，即$\chi^2_{\text{Wald}} = (\hat{\gamma} - \gamma_0)^2/\mathrm{var}\{\hat{\gamma}\} = (\hat{\gamma} - \gamma_0)^2 I_n(\hat{\gamma})$，其服从自由度为 1 的卡方分布。若$\chi^2_{\text{Wald}} > \chi^2_{1-\alpha}(1)$，拒绝$H_0$。

(2) 似然比检验

似然比检验(likelihood ratio test, LRT)是另一个广泛使用的假设检验方法，当对仅有的一个总体参数γ进行假设检验$H_0: \gamma = \gamma_0$时，似然比统计量的定义为：

$$\lambda = \frac{L(\gamma_0 | \boldsymbol{x})}{L(\hat{\gamma} | \boldsymbol{x})} \tag{4}$$

由此可见，似然比统计量为当H_0成立时的似然函数除以最大的似然函数，显然分母中γ的取值便是极大似然估计$\hat{\gamma}$。由于似然函数为非负数，似然比统计量的值域为 0 到 1。当$\hat{\gamma} = \gamma_0$时，似然比统计量达到最大值1，说明有充足的证据支持H_0；反之当似然比统计量接近 0 时，说明在H_0条件下出现该样本的可能性极低，倾向于拒绝H_0。

在大样本下，对数似然比检验统计量收敛于自由度为 1 的卡方分布，即：

$$\chi^2_{\text{LRT}} = -2\log\frac{L(\gamma_0 | \boldsymbol{x})}{L(\hat{\gamma} | \boldsymbol{x})} = 2\left[\log L(\hat{\gamma} | \boldsymbol{x}) - \log L(\gamma_0 | \boldsymbol{x})\right] \xrightarrow{d} \chi^2(1) \tag{5}$$

给定显著性水平α，当$\chi^2_{\text{LRT}} > \chi^2_{1-\alpha}(1)$ 时，拒绝H_0。反转上述检验统计量可以得到γ的轮廓似然(profile likelihood)置信区间[8]。以构造 95%轮廓似然置信区间为例：

$$\begin{aligned} &2\left[\log L(\hat{\gamma} | \boldsymbol{x}) - \log L(\gamma_{\text{upper}} | \boldsymbol{x})\right] \\ &= 2\left[\log L(\hat{\gamma} | \boldsymbol{x}) - \log L(\gamma_{\text{lower}} | \boldsymbol{x})\right] \\ &= \chi^2_{0.95} = 3.84 \end{aligned} \tag{6}$$

求解式(6)中的γ_{upper}和γ_{lower}可得 95%置信区间的上、下限。

(3) 得分检验

得分检验(score test)又被称为拉格朗日乘子检验(Lagrange multiplier test)。定义得分统计

量(score)为对数似然函数的一阶偏导，即：

$$S(\gamma)=\frac{\partial}{\partial\gamma}\log L(\gamma|\boldsymbol{x})=\sum_{i=1}^{n}\frac{\partial}{\partial\gamma}\log f(x_i|\gamma) \tag{7}$$

当$H_0: \gamma=\gamma_0$成立时，$S(\gamma_0)$在大样本下服从均值为 0，方差为$I_n(\gamma_0)$的正态分布。因此，得分检验的检验统计量为：

$$Z_{\text{score}}=\frac{S(\gamma_0)}{\sqrt{\text{var}\{S(\gamma_0)\}}}=\frac{S(\gamma_0)}{\sqrt{I_n(\gamma_0)}} \tag{8}$$

与 Wald 检验类似，在大样本下Z_{score}服从标准正态分布。当以显著性水平α进行双侧检验时，若$|Z_{\text{score}}|>Z_{1-\alpha/2}$，则拒绝$H_0$。对$Z_{\text{score}}$取平方可得到卡方统计量的形式，即$\chi^2_{score}=S(\gamma_0)^2/I_n(\gamma_0)$，当$\chi^2_{score}>\chi^2_{1-\alpha}(1)$时，拒绝$H_0$。

(4) 多元统计中的三种大样本检验

推广到随机变量中含有多个未知总体参数的场景。记$\boldsymbol{\theta}=\{\theta_1,\dots,\theta_p\}^T$是$p$维未知的总体参数向量，参数的极大似然估计$\widehat{\boldsymbol{\theta}}$在大样本下会依分布收敛于均值为$\boldsymbol{\theta}$，方差-协方差矩阵为$\boldsymbol{I_n}^{-1}(\boldsymbol{\theta})$的多元正态分布[9]。$\boldsymbol{I_n}(\boldsymbol{\theta})$此时被称为信息矩阵，具有如下形式：

$$\begin{aligned}\boldsymbol{I}_n(\boldsymbol{\theta})&=E\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log L(\boldsymbol{\theta}|\boldsymbol{x})\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log L(\boldsymbol{\theta}|\boldsymbol{x})\right]^T\right\}\\&=-E\left(\frac{\partial^2\log L(\boldsymbol{\theta}|\boldsymbol{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)\\&=-nE\left(\frac{\partial^2\log f(x_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)\end{aligned} \tag{9}$$

如果对p维的参数向量$\boldsymbol{\theta}$做双侧假设检验：$H_0: \boldsymbol{\theta}=\boldsymbol{\theta_0}, H_1: \boldsymbol{\theta}\neq\boldsymbol{\theta_0}$，三种大样本检验统计量的构造如下：

(1) Wald 检验统计量：$\chi^2_{\text{Wald}}=(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\boldsymbol{I_n}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$；

(2) 似然比检验统计量：$\chi^2_{\text{LRT}}=2[\log L(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})-\log L(\boldsymbol{\theta}_0|\boldsymbol{x})]$；

(3) 得分检验统计量：$\chi^2_{\text{score}}=\boldsymbol{S}(\boldsymbol{\theta}_0)^{\boldsymbol{T}}\boldsymbol{I_n}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{S}(\boldsymbol{\theta}_0)$。

在原假设下，以上三个检验统计量均渐近服从自由度为p的卡方分布。在给定显著性水平α下，当检验统计量大于$\chi^2_{1-\alpha}(p)$时拒绝原假设H_0。

假设只对p个未知总体参数中某一个分量进行假设检验，记$\boldsymbol{\theta}=\binom{\boldsymbol{\beta}}{\gamma}$为参数向量，$\gamma$为待检验第 p 个参数，其余参数$\boldsymbol{\beta}$为冗余参数(nuisance parameter)。对γ进行如下假设检验：$H_0: \gamma=\gamma_{\boldsymbol{0}}, H_1: \gamma\neq\gamma_{\boldsymbol{0}}$，则三种大样本检验统计量的构造如下：

(1) Wald 检验统计量：$\chi^2_{\text{Wald}}=(\hat{\gamma}-\gamma_0)^2/\boldsymbol{I_n}^{-1}(\widehat{\boldsymbol{\theta}})_{pp}$；

(2) 似然比检验统计量：$\chi^2_{\mathrm{LRT}} = 2[\log L(\hat{\boldsymbol{\beta}},\hat{\gamma}|\boldsymbol{x}) - \log L(\hat{\boldsymbol{\beta}}_0,\gamma_{\mathbf{0}}|\boldsymbol{x})]$；

(3) 得分检验统计量：$\chi^2_{\mathrm{score}} = S(\gamma_{\mathbf{0}})^2 \boldsymbol{I_n}^{-\mathbf{1}}(\hat{\boldsymbol{\theta}}_{\mathbf{0}})_{pp}$。

上式中，$\hat{\boldsymbol{\theta}} = \binom{\hat{\beta}}{\hat{\gamma}}$为所有参数的极大似然估计值，而$\hat{\boldsymbol{\theta}}_0 = \binom{\hat{\beta}_0}{\gamma_0}$为$H_0: \gamma = \gamma_0$成立时，冗余参数$\boldsymbol{\beta}$的极大似然估计值。$\boldsymbol{I_n}^{-1}(\hat{\boldsymbol{\theta}})_{pp}$和$1/\boldsymbol{I_n}^{-1}(\hat{\boldsymbol{\theta}}_0)_{pp}$为矩阵对角线上第$p$个元素，分别代表极大似然估计$\hat{\gamma}$和得分统计量$S(\gamma_0)$的方差分量。以上检验统计量大样本下均服从卡方分布，自由度为 1[10]。

表 1 汇总了三种检验统计量的性质。

表 1. 三种渐近检验统计量的性质

	全局检验$H_0: \boldsymbol{\theta} = \boldsymbol{\theta_0}$	局部检验$H_0: \gamma = \gamma_0$	与似然函数的关系
Wald 检验	$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})^T \boldsymbol{I}_n(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) \overset{H_0}{\to} \chi^2_p$	$\frac{(\hat{\gamma} - \gamma_0)^2}{\boldsymbol{I_n}^{-1}(\hat{\boldsymbol{\theta}})_{pp}} \overset{H_0}{\to} \chi^2_1$	只需要优化无约束模型
似然比检验	$2[\log L(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) - \log L(\boldsymbol{\theta_0}|\boldsymbol{x})] \overset{H_0}{\to} \chi^2_p$	$2[\log L(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) - \log L(\hat{\boldsymbol{\theta}}_0|\boldsymbol{x})] \overset{H_0}{\to} \chi^2_1$	需要同时优化有约束和无约束模型
得分检验	$\boldsymbol{S}(\boldsymbol{\theta_0})^T \boldsymbol{I_n}^{-1}(\boldsymbol{\theta_0}) \boldsymbol{S}(\boldsymbol{\theta_0}) \overset{H_0}{\to} \chi^2_p$	$S(\gamma_0)^2 \boldsymbol{I_n}^{-1}(\hat{\boldsymbol{\theta}}_0)_{pp} \overset{H_0}{\to} \chi^2_1$	只需要优化有约束模型

此处，全局检验是指对整个模型的假设检验，即同时检验所有参数$\boldsymbol{\theta}$是否满足原假设$H_0: \boldsymbol{\theta} = \boldsymbol{\theta_0}$，例如检验$\boldsymbol{\theta_0} = \{\theta_1, \dots, \theta_p\}^T = \mathbf{0}$；局部检验是指在控制了其他协变量后，检验某一个参数γ。三种检验都和似然函数密切相关，对所有参数进行极大似然估计的模型被称为无约束模型；而H_0成立时，仅对冗余参数进行极大似然估计的模型被称为有约束模型。由此可见，Wald 只需拟合无约束模型，得分检验只需拟合有约束模型，而似然比检验需同时拟合两个模型。

1.2 logistic 回归模型

GWAS 研究中常用的 logistic 回归模型如下：

$$\mathrm{logit}\,[\Pr(Y_i = 1 | X_i, G_i)] = \mathrm{logit}\ \pi_i = \boldsymbol{X_i}^T \boldsymbol{\beta} + G_i \gamma,\ i = 1, 2, \cdots, n \tag{10}$$

式(10)中，$Y_i = 1$或0，分别代表第i个个体是病例或对照，π_i表示该样本是病例的概率，$\boldsymbol{X_i}$是$k \times 1$维截距项与协变量，G_i代表第i个个体待检验的遗传位点，编码为次等位基因数目，$G_i = 0, 1, 2$。$\boldsymbol{\beta}$是$k \times 1$维协变量的系数向量，γ是位点的效应系数，等于其优势比(odds ratio, OR)的对数。GWAS 关注该遗传位点是否与疾病相关，则原假设$H_0: \gamma = 0$。

回归系数$\boldsymbol{\beta}$和γ均可采用极大似然方法进行估计。记$\boldsymbol{\theta} = \binom{\boldsymbol{\beta}}{\gamma}$为 k+1 维的系数向量，$\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{G}$分别代表了每一行为$y_i, \boldsymbol{X_i}$和$G_i$的$n$行样本矩阵，logistic 回归的似然函数如下：

$$L(\boldsymbol{\theta}\,|\,\boldsymbol{y},\boldsymbol{X},\boldsymbol{G})=\prod_{i=1}^{n}\Pr\,(Y_i=y_i\,|\,\boldsymbol{\theta},\boldsymbol{X_i},G_i)=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{11}$$

对数似然函数的一阶偏导为得分统计量。当H_0成立时，该模型的得分统计量为：

$$\begin{aligned}S(0)&=\sum_{i=1}^{n}\frac{\partial}{\partial\gamma}\log\,\Pr\,(Y_i=y_i\,|\,\boldsymbol{X}_i,\boldsymbol{\beta},G_i,\gamma)\bigg|_{\gamma=0}\\&=\sum_{i=1}^{n}G_i(Y_i-\pi_i)\\&\approx\sum_{i=1}^{n}G_i(Y_i-\hat{\pi}_{0i})\\&=\boldsymbol{G^T R}\end{aligned} \tag{12}$$

当H_0成立时，零模型的π_i可用极大似然估计$\hat{\pi}_{0i}=\exp\left(\boldsymbol{X_i^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}\right)/\left[1+\exp\left(\boldsymbol{X_i^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}\right)\right]$进行替代。由此可见，$\boldsymbol{R}$为零模型的残差向量$\boldsymbol{R}=\{Y_1-\hat{\pi}_{01},\ldots,Y_n-\hat{\pi}_{0n}\}^T$。

对数似然函数的二阶偏导为 Fisher 信息矩阵，即

$$\begin{aligned}\boldsymbol{I_n}(\boldsymbol{\theta})&=-E\left(\frac{\partial^2\log L(\boldsymbol{\theta}\,|\,\boldsymbol{y},\boldsymbol{X},\boldsymbol{G})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right)\\&=-E\left(\sum_{i=1}^{n}\frac{\partial^2\log\,\Pr\,(Y_i=y_i\,|\,\boldsymbol{\theta},\boldsymbol{X_i},G_i)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right)\\&=\begin{bmatrix}\boldsymbol{X^T WX} & \boldsymbol{X^T WG}\\ \boldsymbol{G^T WX} & \boldsymbol{G^T WG}\end{bmatrix}\end{aligned} \tag{13}$$

式(13)中的$\boldsymbol{W}$是Y_i的方差-协方差矩阵，即以$\pi_i(1-\pi_i)$为对角元素的对角矩阵，估计时π_i用相应的极大似然估计$\hat{\pi}_i$替代。通过矩阵运算可以得到$\boldsymbol{\theta}$的 Fisher 信息矩阵的逆矩阵$\boldsymbol{I_n}^{-1}(\boldsymbol{\theta})$，其对角线上的第 k+1 个元素（即参数γ对应元素）为$\boldsymbol{I_n}^{-1}(\boldsymbol{\theta})_{k+1,k+1}=1/[\boldsymbol{G^T WG}-\boldsymbol{G^T WX}(\boldsymbol{X^T WX})^{-1}\boldsymbol{X^T WG}]$。

结合以上推导，logistic 回归的三种大样本检验统计分别由下式给出：

(1) Wald 检验统计量：

$$\begin{aligned}Z_{\text{Wald}}&=\frac{\hat{\gamma}}{\text{se}(\hat{\gamma})}=\frac{\hat{\gamma}}{\sqrt{\boldsymbol{I_n}^{-1}\left(\hat{\boldsymbol{\theta}}\right)_{k+1,k+1}}}\\&=\hat{\gamma}\sqrt{\boldsymbol{G^T\hat{W}G}-\boldsymbol{G^T\hat{W}X}\left(\boldsymbol{X^T\hat{W}X}\right)^{-1}\boldsymbol{X^T\hat{W}G}}\end{aligned} \tag{14}$$

(2) 似然比检验统计量：

$$\begin{aligned}\chi^2_{\text{LRT}}&=2\left[\log L\left(\hat{\boldsymbol{\beta}},\hat{\gamma}\,\middle|\,\boldsymbol{y},\boldsymbol{X},\boldsymbol{G}\right)-\log L\left(\hat{\boldsymbol{\beta}}_0,\gamma_0=0\,\middle|\,\boldsymbol{y},\boldsymbol{X},\boldsymbol{G}\right)\right]\\&=2\left\{\sum_{i=1}^{n}\left[y_i\log\frac{\hat{\pi}_i}{(1-\hat{\pi}_i)}+\log(1-\hat{\pi}_i)\right]-\sum_{i=1}^{n}\left[y_i\log\frac{\hat{\pi}_{0i}}{(1-\hat{\pi}_{0i})}+\log(1-\hat{\pi}_{0i})\right]\right\}\\&=2\sum_{i=1}^{n}\left[y_i\log\frac{\hat{\pi}_i(1-\hat{\pi}_{0i})}{\hat{\pi}_{0i}(1-\hat{\pi}_i)}+\log\frac{1-\hat{\pi}_i}{1-\hat{\pi}_{0i}}\right]\end{aligned} \tag{15}$$

(3) 得分检验统计量：

$$
\begin{aligned}
Z_{\text{score}} &= \frac{S(0)}{\sqrt{\operatorname{var}\{S(0)\}}} = \frac{S(0)}{\sqrt{1/\boldsymbol{I_n}^{-1}\left(\hat{\boldsymbol{\theta}}_0\right)_{k+1,k+1}}} \\
&= \frac{\boldsymbol{G^T R}}{\sqrt{\boldsymbol{G^T}\hat{\boldsymbol{W}}_0\boldsymbol{G} - \boldsymbol{G^T}\hat{\boldsymbol{W}}_0\boldsymbol{X}\left(\boldsymbol{X^T}\hat{\boldsymbol{W}}_0\boldsymbol{X}\right)^{-1}\boldsymbol{X^T}\hat{\boldsymbol{W}}_0\boldsymbol{G}}}
\end{aligned}
\tag{16}
$$

值得注意的是，在 GWAS 无缺失数据的情况下，得分检验只需拟合一次固定协变量组合的空模型，便可得到式(16)中的重要参数$\boldsymbol{R}$和$\boldsymbol{W_0}$，故运算效率很高，在大规模 GWAS 中得到广泛应用[1, 11, 12]。记$\widetilde{\boldsymbol{G}} = \boldsymbol{G} - \boldsymbol{X}\left(\boldsymbol{X^T}\widehat{\boldsymbol{W}}_{\boldsymbol{0}}\boldsymbol{X}\right)^{\boldsymbol{-1}}\boldsymbol{X^T}\widehat{\boldsymbol{W}}_{\boldsymbol{0}}\boldsymbol{G}$，为协变量调整的位点向量，则可计算另一种形式的得分统计量(式 17)及其渐近方差(式 18)。两种方法所得检验统计量Z_{score}完全相等[11]，但是具有更为简洁的形式。

$$
S(0) = \sum_{i=1}^{n} \tilde{G}_i (Y_i - \hat{\pi}_{0i}) = \tilde{\boldsymbol{G}}^{\boldsymbol{T}} \boldsymbol{R} \tag{17}
$$

$$
\operatorname{var}\{S(0)\} = \tilde{\boldsymbol{G}}^{\boldsymbol{T}} \hat{\boldsymbol{W}}_0 \tilde{\boldsymbol{G}} \tag{18}
$$

$$
Z_{\text{score}} = \frac{S(0)}{\sqrt{\operatorname{var}\{S(0)\}}} = \frac{\tilde{\boldsymbol{G}}^{\boldsymbol{T}} \boldsymbol{R}}{\sqrt{\tilde{\boldsymbol{G}}^{\boldsymbol{T}} \hat{\boldsymbol{W}}_0 \tilde{\boldsymbol{G}}}} \tag{19}
$$

因此，GWAS 中 logistic 回归一般采用如下步骤进行得分检验，以提高计算和存储的效率。首先，拟合结局与协变量的 logistic 回归模型，算得如$\boldsymbol{R}$、$\boldsymbol{W_0}$；然后，拟合待检验位点和协变量的线性回归模型，算得$\widetilde{\boldsymbol{G}}$；最后代入式(19)，便可得$Z_{\text{score}}$。若分析不同位点，只需更新$\widetilde{\boldsymbol{G}}$即可，从而避免反复进行极大似然法估计参数，计算速度较快。但是，该方法不适用于构造置信区间。

1.3 Cox 比例风险模型

Cox 比例风险模型可用式(20)表达：

$$
\lambda(t; \boldsymbol{X_i}, G_i) = \lambda_0(t) \exp(\boldsymbol{X_i}^T \boldsymbol{\beta} + G_i \gamma), i = 1, 2, \cdots, n \tag{20}
$$

此处，$\lambda(t; \boldsymbol{X_i}, G_i)$表示样本在$t$时刻的风险函数，$\lambda_0(t)$表示基线风险函数。风险函数$\lambda(t)$代表样本在$t$时刻发生结局事件的瞬时概率。同样，原假设为$H_0: \gamma = 0$。该模型的似然函数由于包含未知的风险函数$\lambda_0$，无法直接进行极大似然估计。因此，Cox 提出采用偏似然函数估计目标参数[13]。令$\mathcal{R}(t) = \{j: T_j \geq t\}$表示在$t$时刻承受风险的样本集合，则似然函数中$\lambda_0$可以消掉，得到偏似然函数：

$$L_p(\boldsymbol{\beta},\gamma|\boldsymbol{\Delta},\boldsymbol{T},\boldsymbol{X},\boldsymbol{G})=\prod_{i=1}^{n}\left[\frac{\exp(\boldsymbol{X_i^T\beta}+G_i\gamma)}{\sum_{j\in\mathcal{R}(t_i)}\exp(\boldsymbol{X_i^T\beta}+G_i\gamma)}\right]^{\Delta_i} \tag{21}$$

观测到的生存结局由两部分组成：①发生结局事件或删失，分别记为$\Delta_i=1$或0；②生存事件T_i。式(21)仅取决于事件时间的秩次，且没有直接使用删失和未删失的事件时间数值，因此被称为偏似然函数(partial likelihood function)。利用偏似然函数可对模型中参数$\boldsymbol{\beta}$和γ进行极大偏似然估计。当生存数据中存在较多打结(ties)数据时，即存在大量相同取值的生存或删失数据，则可采用 Breslow's 近似或 Efron's 近似等方法校正偏似然函数[14]。

偏似然函数的γ一阶偏导为得分统计量。当H_0成立时，得分统计量为：

$$\begin{aligned}S(0)&=\left.\frac{\partial\log L_p(\boldsymbol{\beta},\gamma)}{\partial\gamma}\right|_{\gamma=0}\\&=\sum_{i=1}^{n}\Delta_i\left\{G_i-\frac{\sum_{j\in\mathcal{R}(t_i)}G_j\exp(\boldsymbol{X_j}^T\boldsymbol{\beta})}{\sum_{j\in\mathcal{R}(t_i)}\exp(\boldsymbol{X_j}^T\boldsymbol{\beta})}\right\}\\&=\sum_{i=1}^{n}G_i\left[\Delta_i-H(t_i,\boldsymbol{X},\boldsymbol{\beta})\right]\\&\approx\sum_{i=1}^{n}G_i\left[\Delta_i-H\left(t_i,\boldsymbol{X},\hat{\boldsymbol{\beta}}_0\right)\right]\\&=\sum_{i=1}^{n}G_iR_i=\boldsymbol{G^TR}\end{aligned} \tag{22}$$

此处，$H(t_i,\boldsymbol{X},\boldsymbol{\beta})$是累积风险函数，$R_i=\Delta_i-H\left(t_i,\boldsymbol{X},\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}\right)$为零模型下样本的鞅残差(martingale residual)。可见，Cox 回归模型与 logistic 回归模型的得分统计量具有一致的形式[15]。

对数偏似然函数二阶偏导为 Fisher 信息矩阵。记$\boldsymbol{\theta}=\binom{\boldsymbol{\beta}}{\gamma}$为 k+1 维系数向量，并计算偏似然函数的 Fisher 信息矩阵：

$$\begin{aligned}\boldsymbol{I_n}(\boldsymbol{\theta})&=E\left[-\frac{\partial^2\log L_p(\boldsymbol{\theta}|\boldsymbol{\Delta},\boldsymbol{T},\boldsymbol{X},\boldsymbol{G})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]\\&=\begin{bmatrix}\boldsymbol{X^TQX}&\boldsymbol{X^TQG}\\\boldsymbol{G^TQX}&\boldsymbol{G^TQG}\end{bmatrix}\end{aligned} \tag{23}$$

此处，$\boldsymbol{Q}$是结局事件的方差-协方差矩阵，具体公式为：

$$\begin{aligned}&\boldsymbol{Q}=\sum_{i=1}^{n}\frac{\Delta_i}{w_i(\boldsymbol{\theta})^2}\left[w_i(\boldsymbol{\theta})\operatorname{diag}\{\boldsymbol{\Lambda_i}\}-\boldsymbol{\Lambda_i}\boldsymbol{\Lambda_i}^T\right]\\&\boldsymbol{\Lambda_i}=\boldsymbol{1}_{\mathcal{R}(t_i)}*\exp(\boldsymbol{X\beta}+\boldsymbol{G}\gamma)\\&w_i(\boldsymbol{\theta})=\boldsymbol{1}_{\mathcal{R}(t_i)}^T\exp(\boldsymbol{X\beta}+\boldsymbol{G}\gamma)=\sum\boldsymbol{\Lambda_i}\end{aligned} \tag{24}$$

式中，$\boldsymbol{1}_{\mathcal{R}(t_i)}$为一个$n$维的示性向量，用于指示生存时间长于第$i$个个体的样本$j$，即

$\{j: T_j \geq t_i\}$记为 1，反之记为 0。diag{·}表示以该向量为对角线元素的对角矩阵。$*$为哈达马积，定义为两个矩阵对应元素的乘积，故$w_i(\boldsymbol{\theta})$为向量$\boldsymbol{\Lambda_i}$中每一个元素之和。

于是，三大渐近检验统计量的形式如下：

(1) Wald 检验统计量：

$$\begin{aligned} Z_{\text{Wald}} &= \frac{\hat{\gamma}}{\text{se}(\hat{\gamma})} = \frac{\hat{\gamma}}{\sqrt{\boldsymbol{I_n}^{-1}\left(\hat{\boldsymbol{\theta}}\right)_{k+1,k+1}}} \\ &= \hat{\gamma}\sqrt{\boldsymbol{G^T}\hat{\boldsymbol{Q}}\boldsymbol{G} - \boldsymbol{G^T}\hat{\boldsymbol{Q}}\boldsymbol{X}\left(\boldsymbol{X^T}\hat{\boldsymbol{Q}}\boldsymbol{X}\right)^{-1}\boldsymbol{X^T}\hat{\boldsymbol{Q}}\boldsymbol{G}} \end{aligned} \tag{25}$$

(2) 似然比检验统计量：

$$\begin{aligned} \chi^2_{\text{LRT}} &= 2\left[\log L_p\left(\hat{\boldsymbol{\beta}}, \hat{\gamma} \middle| \boldsymbol{\Delta}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{G}\right) - \log L_p\left(\hat{\boldsymbol{\beta}}_0, \gamma_0 = 0 \middle| \boldsymbol{\Delta}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{G}\right)\right] \\ &= 2\sum_{i=1}^{n}\Delta_i\left[\boldsymbol{X_i^T}\hat{\boldsymbol{\beta}} + G_i\hat{\gamma} - \boldsymbol{X_i^T}\hat{\boldsymbol{\beta}}_0 - \log\left(\frac{\sum\limits_{j\in\mathcal{R}(t_i)}\exp\left(\boldsymbol{X_i^T}\hat{\boldsymbol{\beta}} + G_i\hat{\gamma}\right)}{\sum\limits_{j\in\mathcal{R}(t_i)}\exp\left(\boldsymbol{X_i^T}\hat{\boldsymbol{\beta}}_0\right)}\right)\right] \\ &= 2\sum_{i=1}^{n}\Delta_i\left[\boldsymbol{X_i^T}\hat{\boldsymbol{\beta}} + G_i\hat{\gamma} - \boldsymbol{X_i^T}\hat{\boldsymbol{\beta}}_0 - \log\left(\frac{w_i\left(\hat{\boldsymbol{\theta}}\right)}{w_i\left(\hat{\boldsymbol{\theta}}_0\right)}\right)\right] \end{aligned} \tag{26}$$

(3) 得分检验统计量(令$\widetilde{\boldsymbol{G}} = \boldsymbol{G} - \boldsymbol{X}\left(\boldsymbol{X^T}\widehat{\boldsymbol{Q}}_{\boldsymbol{0}}\boldsymbol{X}\right)^{-\boldsymbol{1}}\boldsymbol{X^T}\widehat{\boldsymbol{Q}}_{\boldsymbol{0}}\boldsymbol{G}$)：

$$\begin{aligned} Z_{\text{score}} &= \frac{S(0)}{\sqrt{\text{var}\{S(0)\}}} = \frac{S(0)}{\sqrt{1/\boldsymbol{I_n}^{-1}\left(\hat{\boldsymbol{\theta}}_0\right)_{k+1,k+1}}} \\ &= \frac{\boldsymbol{G^T}\boldsymbol{R}}{\sqrt{\boldsymbol{G^T}\hat{\boldsymbol{Q}}_0\boldsymbol{G} - \boldsymbol{G^T}\hat{\boldsymbol{Q}}_0\boldsymbol{X}\left(\boldsymbol{X^T}\hat{\boldsymbol{Q}}_0\boldsymbol{X}\right)^{-1}\boldsymbol{X^T}\hat{\boldsymbol{Q}}_0\boldsymbol{G}}} \\ &= \frac{\tilde{\boldsymbol{G}}^{\boldsymbol{T}}\boldsymbol{R}}{\sqrt{\tilde{\boldsymbol{G}}^{\boldsymbol{T}}\hat{\boldsymbol{Q}}_0\tilde{\boldsymbol{G}}}} \end{aligned} \tag{27}$$

1.3 混合效应模型

种群结构和亲缘关系是影响 GWAS 结果的重要混杂因素。混合效应模型是常用的解决方案[1, 5, 16-18]。鉴于 Cox 混合效应模型也拥有相似的形式，本文以 logistic 混合效应模型作为示例，具体模型如下：

$$\text{logit}\left[Pr(Y_i = 1 | X_i, G_i)\right] = \text{logit}\ \pi_i = \boldsymbol{X_i}^{\boldsymbol{T}}\boldsymbol{\beta} + G_i\gamma + \xi_i,\ i = 1, 2, \cdots, n \tag{28}$$

与传统 logistic 回归模型不同，式(28)增加了用于刻画种群结构和亲缘关系的ξ_i。$\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_n\}^T$为随机效应，服从多元正态分布$N_n(0, \tau\boldsymbol{V})$。其中，$\boldsymbol{V}$为$n \times n$的遗传关系矩阵(genomic relationship matrix, GRM)，τ是用于刻画方差分量的尺度参数。该模型中的$\boldsymbol{\beta}$、γ和

τ三个参数可使用惩罚准似然(penalized qusai-maximum likelihood, PQML)[19]或平均信息限制性最大似然(average information restricted maximum likelihood, AI-REML)算法[20]进行估计。在 H_0 条件下，得分统计量为：$S(0)=\sum_{i=1}^{n}G_iR_i=\sum_{i=1}^{n}G_i(Y_i-\hat{\pi}_{0i})$，其方差为：$\mathrm{var}\{S(0)\}=\boldsymbol{G^T\hat{\Sigma}_0G}-\boldsymbol{G^T\hat{\Sigma}_0X}\left(\boldsymbol{X^T\hat{\Sigma}_0X}\right)^{-1}\boldsymbol{X^T\hat{\Sigma}_0G}$。$\hat{\pi}_{0i}$是$H_0$条件下估计的事件发生率。$\boldsymbol{\hat{\Sigma}_0}=\left[\boldsymbol{\widehat{W}_0}^{-1}+\hat{\tau}\boldsymbol{V}\right]^{-1}=[\mathrm{diag}\{\hat{\pi}_{0i}/(1-\hat{\pi}_{0i})\}^{-1}+\hat{\tau}\boldsymbol{V}]^{-1}$。由此可见，数据的总变异来自于个体变异和 GRM。同样，得分检验统计量Z_{score}收敛到标准正态分布。

对于生存分析的混合效应模型，将$R_i=(Y_i-\hat{\pi}_{0i})$替换为鞅残差，用前文中的$\boldsymbol{Q_0}$替代$\boldsymbol{W_0}$，便可进行得分检验。

二、极端不平衡数据对经典统计量分布的影响

通过模拟试验结果阐述极端不平衡数据对经典统计量分布的影响。简单起见，模型暂不考虑协变量，假定位点和结局存在如下关系：

$$\mathrm{logit}\left[\Pr(Y_i=1|X_i,G_i)\right]=a+G_i\times\beta,\ i=1,2,\cdots,n \tag{29}$$

参考 UKB 中的肺癌数据设置模拟情景，即假设 300,000 人的队列中肺癌基线患病率为1%，约 3,000 例病例，故截距项$a=-4.595$。待检验位点G采用相加模型编码，$G_i=0,1,2$。分别设置两种情景：位点G的 MAF 分别为 0.01、0.001，分别代表低频、罕见变异。令$\beta=0$ 生成模拟样本，分别用 Wald 检验、似然比检验和得分检验对位点G的系数进行假设检验。模拟试验重复 100 万次，获得各个统计量的经验分布，并阴影部分表示上、下 2.5%分位数。如图 1 所示，当 MAF 取 0.01 时，检验统计量的经验分布与对应的标准正态分布或 χ^2 分布较为接近。双侧检验 P 值的 QQ 图显示，基因组膨胀因子 (genomic inflation factor) λ趋近于1。如图 2 所示，当 MAF 下降到 0.001 时，即便样本量已达到 30 万人，检验统计量的经验分布严重偏离标准正态分布或 χ^2 分布——由于自变量和样本量的取值为离散型变量且事件数极少，检验统计量也呈现出明显的离散化趋势。此外，基因组膨胀因子大于 1，也表明可能存在严重的假阳性。

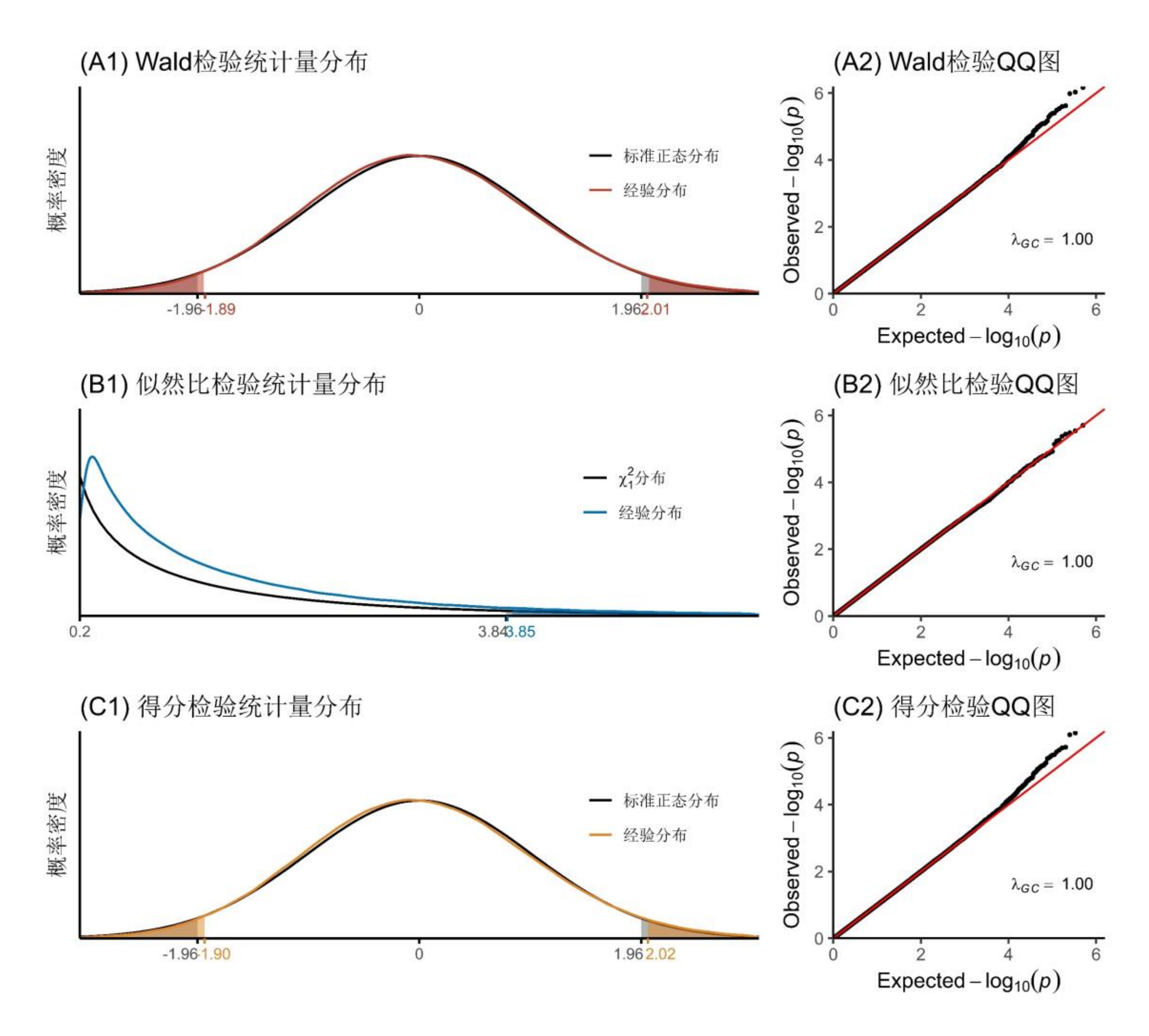


图 1. 轻度不平衡数据三种检验统计量的经验分布

（情境 1：样本量$n = 300{,}000$，患病率$\pi = 1\%$，位点$MAF = 1\%$）

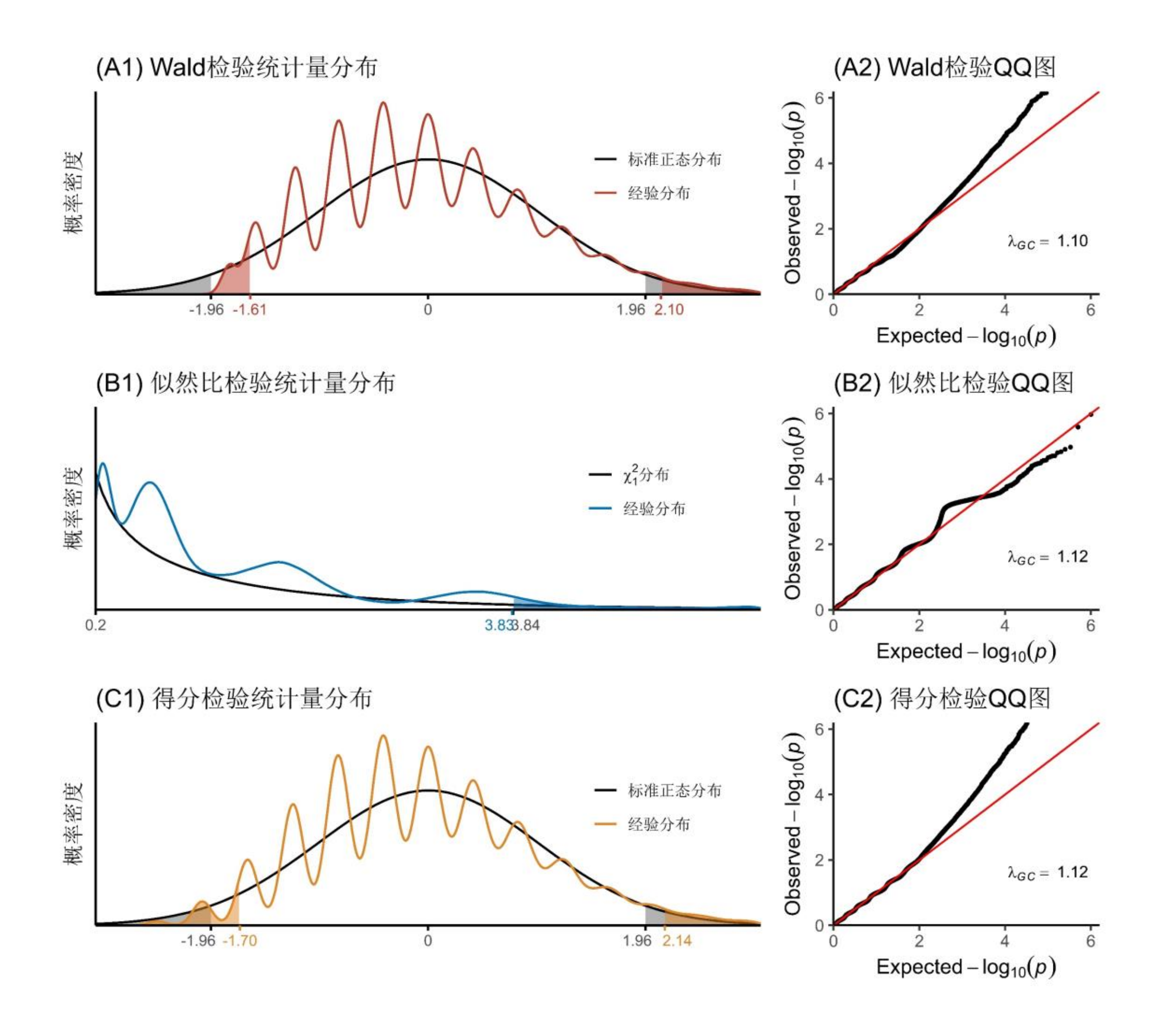


图 2. 极度不平衡数据三种检验统计量的经验分布

（情境 2：样本量$n = 300{,}000$，患病率$\pi = 1\%$，位点$MAF = 0.1\%$）

如果错误地以标准正态分布或 χ^2 分布的界值作为假设检验统计量的界值，那么三种检验无法严格控制一类错误，如表 2 所示。一类错误发生膨胀的程度与检验水准的取值相关。当检验水准设为 0.05 时，在 30 万人的样本下，MAF 无论为 0.01 还是 0.001，双侧检验的第一类错误尚且接近 0.05。然而在 GWAS 研究中，因为需要分析多个位点，所以进行一类错误多重校正。以分析 1000 个位点为例，Bonferroni 方法校正后的检验水准设为5×10^{-5}时，此时在 MAF=0.01 的情景下，Wald 和得分检验一类错误分别膨胀至 6.60×10^{-5} (132%)和 7.22×10^{-5} (144%)，似然比检验的一类错误控制较好；而在 MAF=0.001 的情景下，Wald 和得分检验一类错误更是膨胀至 2.23×10^{-4} (446%)和 7.22×10^{-5} (702%)，似然比检验一类错误控制过于严格，偏保守。若位点 MAF 更小或基线患病率更低，上述现象更明显[4, 11, 21]。主要原因是，GWAS 试图挑选 P 值特别小的位点，极端不平衡数据分析所得三种统计量的经验分布与标准正态分布或 χ^2 分布在分布尾部的差异相对而言较大，这将造成严重的一类错误

膨胀或一类错误保守导致的检验效能低下。此外，由于极度不平衡数据的检验统计量具有偏态的特点，因此单侧检验的一类错误呈现更为严重的偏倚。

表 2. 极端不平衡数据的三种方法假设检验第一类错误率

显著性水平α	MAF	方法	双侧检验 (显著性水平α)		左侧检验 (显著性水平$\alpha/2$)		右侧检验 (显著性水平$\alpha/2$)	
			拒绝域	第一类错误	拒绝域	第一类错误	拒绝域	第一类错误
0.05	0.01	Wald 检验	$\lvert Z_{\text{Wald}}\rvert > 1.96$	4.88×10^{-2}	$Z_{\text{Wald}} < -1.96$	2.08×10^{-4}	$Z_{\text{Wald}} > 1.96$	2.81×10^{-2}
0.05	0.01	似然比检验	$\chi^2_{LRT} > 3.84$	5.04×10^{-2}	—	—	—	—
0.05	0.01	得分检验	$\lvert Z_{\text{score}}\rvert > 1.96$	4.95×10^{-2}	$Z_{\text{score}} < -1.96$	2.12×10^{-2}	$Z_{\text{score}} > 1.96$	2.83×10^{-2}
0.05	0.001	Wald 检验	$\lvert Z_{\text{Wald}}\rvert > 1.96$	3.34×10^{-2}	$Z_{\text{Wald}} < -1.96$	$< 1.00\times10^{-6}$	$Z_{\text{Wald}} > 1.96$	3.34×10^{-2}
0.05	0.001	似然比检验	$\chi^2_{LRT} > 3.84$	4.95×10^{-2}	—	—	—	—
0.05	0.001	得分检验	$\lvert Z_{\text{score}}\rvert > 1.96$	5.11×10^{-2}	$Z_{\text{score}} < -1.96$	1.59×10^{-2}	$Z_{\text{score}} > 1.96$	3.52×10^{-2}
$\mathbf{5\times10^{-5}}$	0.01	Wald 检验	$\lvert Z_{\text{Wald}}\rvert > 4.06$	$\mathbf{6.60\times10^{-5}}$	$Z_{\text{Wald}} < -4.06$	4.00×10^{-6}	$Z_{\text{Wald}} > 4.06$	6.20×10^{-4}
$\mathbf{5\times10^{-5}}$	0.01	似然比检验	$\chi^2_{LRT} > 16.45$	$\mathbf{5.30\times10^{-5}}$	—	—	—	—
$\mathbf{5\times10^{-5}}$	0.01	得分检验	$\lvert Z_{\text{score}}\rvert > 4.06$	$\mathbf{7.20\times10^{-5}}$	$Z_{\text{score}} < -4.06$	4.00×10^{-6}	$Z_{\text{score}} > 4.06$	6.80×10^{-4}
$\mathbf{5\times10^{-5}}$	0.01	Wald 检验	$\lvert Z_{\text{Wald}}\rvert > 4.06$	$\mathbf{2.23\times10^{-4}}$	$Z_{\text{Wald}} < -4.06$	$< 1.00\times10^{-6}$	$Z_{\text{Wald}} > 4.06$	2.23×10^{-4}
$\mathbf{5\times10^{-5}}$	0.001	似然比检验	$\chi^2_{LRT} > 16.45$	$\mathbf{2.70\times10^{-5}}$	—	—	—	—
$\mathbf{5\times10^{-5}}$	0.001	得分检验	$\lvert Z_{\text{score}}\rvert > 4.06$	$\mathbf{3.51\times10^{-4}}$	$Z_{\text{score}} < -4.06$	$< 1.00\times10^{-6}$	$Z_{\text{score}} > 4.06$	3.51×10^{-4}

三、GWAS 研究中极端不平衡数据的常用校正方法

3.1 Firth 校正方法

虽然在大样本下，参数的极大似然估计$\hat{\boldsymbol{\theta}}$会收敛于真值$\boldsymbol{\theta}$，但是当样本量较小或数据极端不平衡时，$\hat{\boldsymbol{\theta}}$作为$\boldsymbol{\theta}$的估计往往是有偏的，偏倚程度取决于得分函数$\boldsymbol{S}(\boldsymbol{\theta})$的曲率[22]。为了消除$O(n^{-1})$阶偏倚，Firth 提出采用惩罚似然函数(式 30)进行参数估计，同时试图控制一类错误[22-24]：

$$\boldsymbol{L}^*(\boldsymbol{\theta}) = \boldsymbol{L}(\boldsymbol{\theta}) \times |\boldsymbol{I}_n(\boldsymbol{\theta})|^{0.5} \tag{30}$$

式中，似然函数的惩罚项$|I_n(\theta)|^{0.5}$是信息矩阵行列式的开方，又被称为 Jeffreys 先验。对惩罚似然函数求一阶偏导可得 Firth 校正的得分统计量为：

$$S^*(\theta_k) = S(\theta) + \frac{1}{2} tr\{\boldsymbol{I}_n(\boldsymbol{\theta})^{-1}[\partial I_n(\theta)/\partial\theta_k]\} \tag{31}$$

θ_k指的是参数向量$\boldsymbol{\theta}$中的第k个分量，也即前文中的γ。式(10)中的 logistic 回归模型，相应的 Firth 校正得分统计量为：

$$S^*(\gamma) = \sum_{i=1}^{n} G_i(Y_i - \pi_i + h_i(1/2 - \pi_i)) \tag{32}$$

此处，h_i是帽子矩阵$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}$的第$i$个对角元素，$\boldsymbol{W}$是$Y_i$的方差-协方差矩阵。从公式(32)可见，Firth 校正同时从因变量和自变量两个角度出发校正得分统计量。从因变量的角度，当π_0越偏离 0.5，Firth 校正力度越大；从自变量的角度，h_i反映自变量的第 i 个观测值与自变量平均观测值间的标化距离，最远为 1，最近为 0。当h_i越接近 1，

Firth 校正力度越大。当因变量和自变量同时为极端不平衡数据时，Firth 法同时从两个角度校正，校正力度较大。

(1) Firth 校正的 Wald 检验和得分检验

极大化式(30)可得 Firth 校正的参数估计$\hat{\theta}^*$。$\hat{\theta}^*$的方差仍用$\boldsymbol{I_n}^{-1}(\hat{\boldsymbol{\theta}}^*)$的对角线元素进行估计，因为其是$[-\partial^2 \log L^*/(\partial\beta)^2]^{-1}$的一阶近似。据此，可算得 Wald 检验统计量，同理也可以用 Firth 校正得分统计量和$\boldsymbol{I_n}^{-1}(\hat{\boldsymbol{\theta}}_{\boldsymbol{0}}^{*})$进行得分检验。但是由于数据极端不平衡的情景下统计量呈现非对称分布，Firth 校正的 Wald 检验以及得分检验仍无法有效控制第一类错误[25]。

(2) Firth 校正的似然比检验

Firth 校正最常用的惩罚似然比检验，该方法也可构造轮廓似然置信区间。该检验统计量是将似然比检验中的似然函数替换为惩罚似然，即：

$$\begin{aligned}\chi^2_{\mathrm{PLRT}} &= 2\left[\log L^*\left(\hat{\boldsymbol{\beta}}^*, \hat{\gamma}^* \middle| \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{G}\right) - \log L^*\left(\hat{\boldsymbol{\beta}}_0^*, \gamma_0 = 0 \middle| \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{G}\right)\right] \\ &= 2\left[\log L\left(\hat{\boldsymbol{\beta}}^*, \hat{\gamma}^* \middle| \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{G}\right) + \log |I_n(\boldsymbol{\beta}, \gamma)|^{0.5}_{\beta=\hat{\beta}^*, \gamma=\hat{\gamma}^*} - \log L\left(\hat{\boldsymbol{\beta}}_0^*, \gamma_0 = 0 \middle| \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{G}\right) - \log |I_n(\boldsymbol{\beta}_0)|^{0.5}_{\beta_0=\hat{\beta}_0^*}\right]\end{aligned} \quad (33)$$

上式中，$\hat{\boldsymbol{\beta}}^*$和$\hat{\gamma}^*$是无约束模型的极大惩罚似然估计，而$\hat{\beta}_0^*$为有约束模型的极大惩罚似然估计。相似地，轮廓似然的 95%置信区间可以通过求解$\log L^*(\hat{\boldsymbol{\beta}}_{\boldsymbol{0}}^*, \gamma_0) \geq \log L^*(\hat{\boldsymbol{\beta}}^*, \hat{\gamma}^*) - 3.84/2$中的$\gamma_0$得到。鉴于 Firth 校正的似然比检验需要通过迭代得到有约束模型和无约束模型的惩罚似然函数极大值，故运算速度较慢。

3.2 鞍点近似方法

鞍点近似方法通过数据从提取分布的信息，构造检验统计量的经验分布，从而根据H_0假定下的经验分布构造 P 值。定义$M(t) = \mathrm{E}(e^{tX}) = \int_{-\infty}^{+\infty} e^{tx} f(x)\, dx$为矩母函数(moment generating function, MGF)，$K(t) = \log(M(t))$为累积量母函数(cumulative generating function, CGF)。累积量母函数具有如下性质，即累积量母函数在 0 位置的一阶、二阶导数分别对应总体均值和方差：

$$\begin{aligned}\mathrm{E}(X) &= K'(0) = \left.\frac{\partial K(t)}{\partial t}\right|_{t=0} \\ \mathrm{var}(X) &= K''(0) = \left.\frac{\partial^2 K(t)}{\partial t^2}\right|_{t=0}\end{aligned} \quad (34)$$

对于任意存在矩母函数的变量分布，其累积分布函数可用鞍点近似方法实现高精度近似[26]：

$$F(x)=\mathrm{P}(X\leq x)\approx\begin{cases}\Phi\left\{\hat{\omega}+\frac{1}{\hat{\omega}}\cdot log\left(\frac{\hat{u}}{\hat{\omega}}\right)\right\} & for\ x\neq\hat{u}\\ \frac{1}{2}+\frac{K'''(0)}{6\sqrt{2\pi}K''(0)^{3/2}} & for\ x=\hat{u}\end{cases} \tag{35}$$

其中，$\widehat{w}=sgn(\hat{t})\sqrt{2\left(\hat{t}x-K(\hat{t})\right)}$，$\hat{\mu}=\hat{t}\sqrt{K''(\hat{t})}$，$sgn(\cdot)$是符号函数，而$\hat{t}$是$K'(\hat{t})=x$的解。正态近似相当于只用了前两个累积量：均值与方差。鞍点近似则使用了整个累积量母函数的信息，其相对误差界限为$O\left(n^{-3/2}\right)$，具有更高的精度。

对于检验统计量S，双侧检验的 P 值是H_0假定下比当前统计量s更极端统计量的比例，即$P=P(S>|s|)+P(S<-|s|)=1-F(|s|)+F(-|s|)$。极端不平衡数据的统计量往往并非正态分布。因此，基于鞍点近似方法所得累积分布函数 $F(x)$所得 P 值到比正态近似法更加精确。

三大检验统计量中得分检验统计量的形式最为简单，因此也最适用于提供鞍点近似构造近似分布。以 logistic 回归模型的H_0成立时的得分统计量$S(0)=\sum_{i=1}^{n}G_i(Y_i-\pi_i)\approx\sum_{i=1}^{n}G_i(Y_i-\hat{\pi}_{0i})$为例，式中只有$Y_i$是随机变量，服从伯努利分布($n=1$的二项分布)的，即得分统计量$S(0)$是$Y_i$的线性组合。若随机变量$X$服从二项分布$Bi(n,\pi)$，其具有显式矩母函数$M_X(t)=(\pi e^t+1-\pi)^n$，而独立随机变量线性组合的矩母函数具有以下公式：

$$\begin{gathered}Y=\sum_{i=1}^{n}a_iX_i+b_i\\ M_Y=\prod_{i=1}^{n}M_{aX+b}(t)=\prod_{i=1}^{n}e^{b_it}M_{X_i}(a_it)\end{gathered} \tag{36}$$

结合以上公式得到得分统计量$S(0)$的累积量母函数及其一、二阶导数。此处将π替换为极大似然估计$\hat{\pi}$即可得到所有 logistic 回归鞍点近似方法需要的统计量。

当结局事件的具体分布未知时，可以借助经验分布求解鞍点近似中所需的统计量。记$\widehat{M}_X(t)=\frac{1}{n}\sum_{i=1}^{n}e^{tx_i}$为经验矩母函数，通过求导运算可估计经验累积量母函数[27]，见公式(37)。

$$
\begin{cases}
\hat{K}_X(t) = \log \hat{M}_X(t) = \log\left(\dfrac{1}{n}\sum_{i=1}^{n} e^{tx_i}\right) \\
\hat{K}'_X(t) = \dfrac{\hat{M}'_X(t)}{\hat{M}_X(t)} = \dfrac{\sum_{i=1}^{n} x_i e^{tx_i}}{\sum_{i=1}^{n} e^{tx_i}} \\
\hat{K}''_X(t) = \dfrac{\hat{M}''_X(t)\hat{M}_X(t) - [\hat{M}'_X(t)]^2}{[\hat{M}_X(t)]^2} = \dfrac{\sum x_i^2 e^{tx_i} \sum e^{tx_i} - [\sum x_i e^{tx_i}]^2}{[\sum e^{tx_i}]^2}
\end{cases}
\tag{37}
$$

图 3 比较了得分统计量的正态近似和鞍点近似累积分布函数，可见鞍点近似方法可以更好地拟合分布的头尾部分。但是，正态近似方法在$x = 0$（相当于鞍点近似方法$x = \hat{u}$）附近的拟合性能并不差。鉴于正态近似方法运算效率远高于鞍点近似方法，因此一般推荐先采用正态近似法计算 P 值，再对特别小的 P 值采用鞍点近似法获得更精确的数值。除计算速度慢外，鞍点近似方法不适合估计效应系数$\hat{\gamma}$及其标准误。

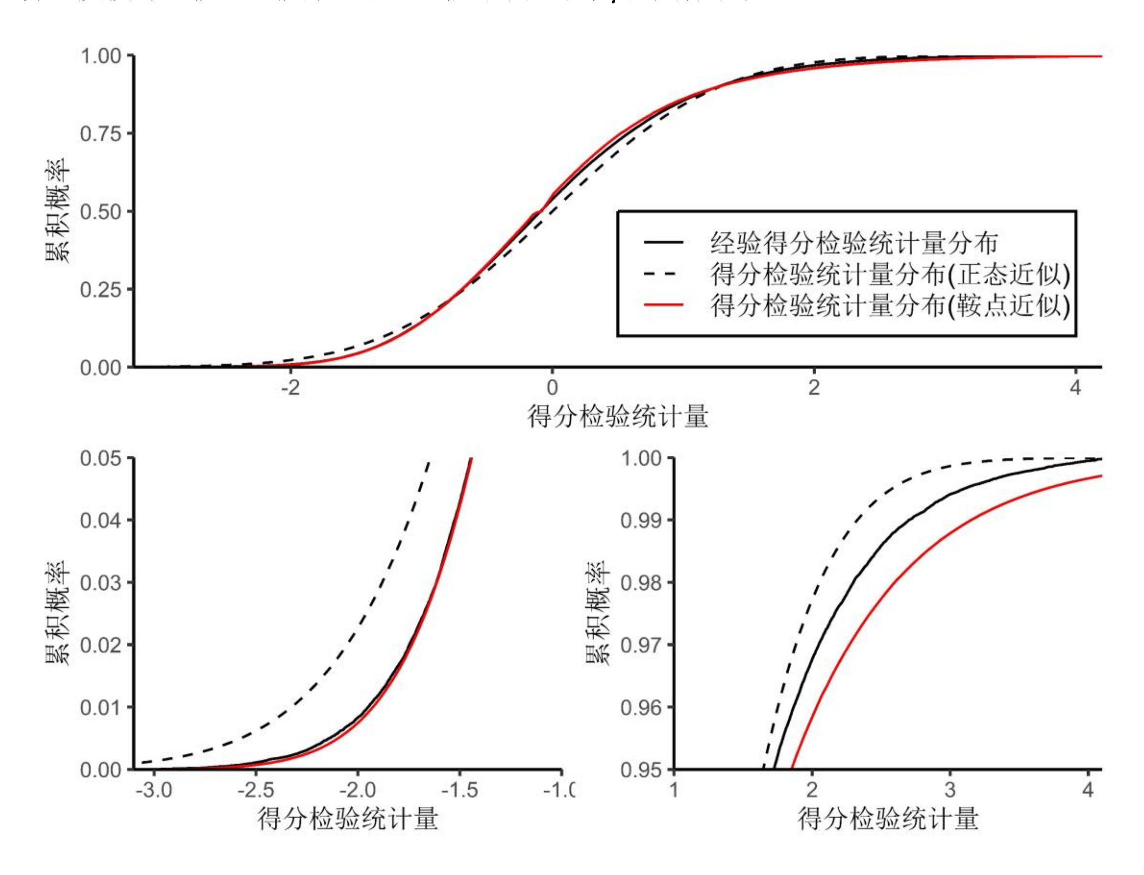


图 3. 得分统计量正态近似和鞍点近似累积分布函数的比较

3.3 总结

本节简述了 Firth 校正方法和鞍点近似方法控制一类错误的核心思想：前者通过校正极大似然估计的偏倚从而提高参数估计的精度和假设检验的性能；后者则通过数据构造检验统计量的经验分布以进行假设检验。两种法均以计算复杂度为代价优化假设检验的统计学

性质。因此，一般推荐分析极端不平衡数据或经典检验 P 值较小可能影响一类错误时使用，避免 GWAS 额外增加计算负担。以模拟试验情景 2 为例，仅对 P 值<0.01 的位点分别实施 Firth 校正方法和鞍点近似方法。如图 4 中所示，Firth 校正的 Wald 得分检验和得分检验仍然出现了一类错误膨胀，Firth 校正的似然比检验稍显保守，这可能是 Firth 校正未能在 GWAS 方法中广泛应用的原因之一。而应用最为广泛的鞍点近似校正的得分检验能够较好地控制一类错误。此外，得益于罕见变异基因型的编码大部分为 0，这部分样本信息对得分统计量无贡献，通过优化算法还可以进一步提高得分检验的计算速度。精确检验、logF 校正、bootstrap 自助法等其他校正检验存在计算效率低下的缺陷，在 GWAS 中应用更少，此处不作介绍，详见相关文献[28, 29]。

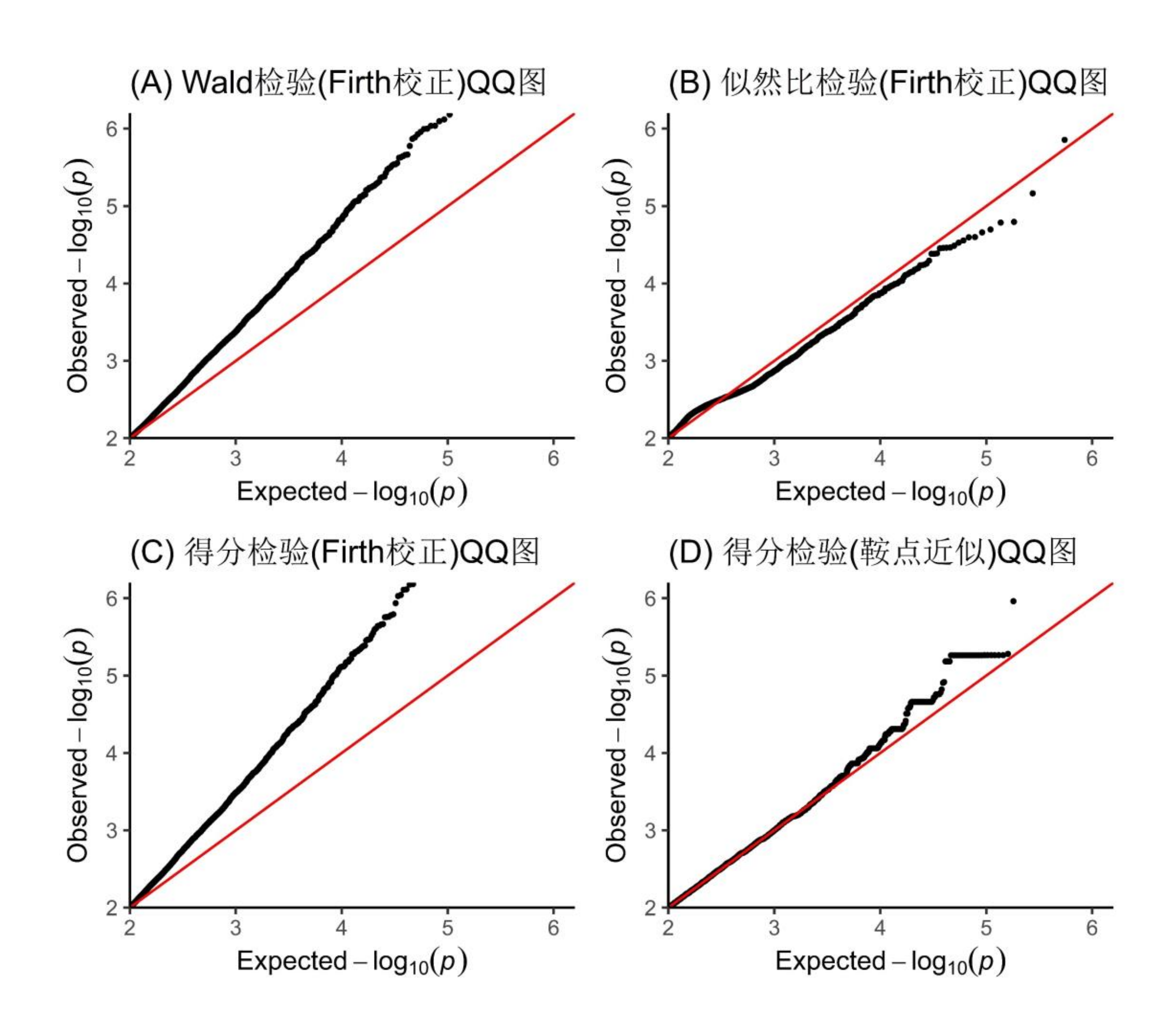


图 4. Firth 校正和鞍点近似方法的 QQ 图

四、基因组学极端不平衡数据的常用软件简介

本节总结了 GWAS 中常用的极端不平衡数据分析软件(表 3)。

表 3. GWAS 常用极端不平衡数据分析软件

软件	结局类型	*P* 值校正方法	亲缘关系校正
R 软件包 *logistf*	二分类结局	Firth	×
R 软件包 *coxhf*	生存结局	Firth	×
PLINK2	二分类结局	Firth	×
R 软件包 *SPAtest*	二分类结局	鞍点近似	×
R 软件包 *SPACox*	生存结局	鞍点近似	×
SAIGE	二分类结局	鞍点近似	√
GATE	生存结局	鞍点近似	√
REGENIE	连续性、二分类结局	Firth、鞍点近似	√

4.1 R 软件包 *logistf* 和 *coxphf*。R 软件中的 *logistf* 和 *coxphf* 分别提供 Firth 校正的 logistic 回归和 Cox 回归，并采用 Wald 统计量和惩罚似然比统计量实现假设检验。虽然 Firth 方法校正了回归模型的系数估计，但使得预测概率向 0.5 偏倚。因此 *logistf* 包中提供了 FLAC 和 FLIC 两种方法来改善预测结果[30]。

4.2 PLINK2。PLINK2 是知名 GWAS 数据分析与处理软件的最新版本[30]。该软件默认处理策略是：先对所有位点进行 logistic 回归，再对模型不收敛的位点采用 Firth 校正 logistic 回归。该软件不能自动鉴别极端不平衡数据，推荐对 MAC < 400 的位点实施 Firth 校正[4]。

4.3 SPAtest。Dey 等[11]将鞍点近似方法引入二分类结局的 GWAS，并发布了 R 软件 *SPAtest* 包。该包对得分统计量位于均值±2 倍标准差范围以内的检验统计量采用正态近似，范围以外的检验统计量采用鞍点近似，以得到假设检验的 *P* 值。此外，用户可以依据 Berry-Esseen 定理的阈值进行判别。研究者还提供了 FastSPA 的版本，针对基因型中大量的 0 数据进行了算法优化，以提高运算效率。

4.4 SPACox。该方法是鞍点近似校正的 Cox 比例风险模型，适合分析大规模生物标本库级别的生存数据[21]。对于二分类结局，得分统计量服从二项分布，易于构造累积量母函数。对于生存数据，鞅残差较难直接得到分布函数。因此，SPACox 直接估计经验累积量母函数，从而构造统计量的经验分布以进行假设检验。此算法可通过 R 软件 *SPACox* 包实现。

4.5 SAIGE。SAIGE 实现了鞍点近似校正的广义混合效应模型，相较于 *SPAtest* 包中的 FastSPA 方法，该方法利用 GRM 额外校正了种群结构和亲缘关系[1]。在运算方面，得分统计量的方差 $varS(0) = \boldsymbol{G^T}\boldsymbol{\hat{\Sigma}_0}\boldsymbol{G} - \boldsymbol{G^T}\boldsymbol{\hat{\Sigma}_0}\boldsymbol{X}\left(\boldsymbol{X^T}\boldsymbol{\hat{\Sigma}_0}\boldsymbol{X}\right)^{-1}\boldsymbol{X^T}\boldsymbol{\hat{\Sigma}_0}\boldsymbol{G} = \boldsymbol{\tilde{G}^T}\boldsymbol{\hat{\Sigma}_0}\boldsymbol{\tilde{G}}$ 较难运算。因此，

SAIGE 首先用部分位点估计变异的比例$\hat{r}=\frac{\widetilde{\boldsymbol{G}}^T\widehat{\boldsymbol{\Sigma}}_0\widetilde{\boldsymbol{G}}}{\widetilde{\boldsymbol{G}}^T\widehat{\boldsymbol{W}}_0\widetilde{\boldsymbol{G}}}$，然后以$\hat{r}\widetilde{\boldsymbol{G}}^T\widehat{\boldsymbol{W}}_0\widetilde{\boldsymbol{G}}$作为$S(0)$的方差，最后通过正态近似或鞍点近似计算 P 值，以同时优化计算速度和统计学性能。

4.6 GATE。GATE 针对生存数据利用修正的泊松对数线性混合效应模型构建似然函数，以应对严重的删失率，并对得分统计量进行鞍点近似[3]。该方法也能够校正种群结构和亲缘关系。整体框架与 SAIGE 类似。

4.7 REGENIE。REGENIE 能够分析二类分结局的极端不平衡数据，校正方法包括：近似 Firth 校正和鞍点近似[31]。相较于 SAIGE，REGENIE 不再估计 GRM，而是采用机器学习的方式构造种群结构方差，从而提高运算效率。为了得到快速的 Firth 校正结果，该方法将协变量效应$\hat{\beta}$作为固定值进行统计量构造，从而避免建立联合分布，以提高运算效率。

五、讨论

大型生物样本库和全基因组测序数据的共享促进了 GWAS 数据的深度挖掘，也推动了遗传统计方法的发展。本文主要介绍了当前 GWAS 中最常用的两种处理极端不平衡数据的校正方法：Firth 校正方法和鞍点近似方法，并总结了相关统计软件，旨在为合理分析 GWAS 数据提供方法和应用参考。

从计算速度来看，为实现大型生物样本库级别的快速分析，多数统计方法采用得分检验进行统计推断，并采用校正 P 值的方法实现第一类错误的控制。从统计性能来看，鞍点近似方法能够较好地控制一类错误，而 Firth 校正方法不尽如人意。相比而言，参数估计仍然处于 GWAS 分析的核心地位。不过，研究者需要注意以下三点：(1) SAIGE 等采用鞍点近似的方法，只调整了假设检验的 P 值，未校正参数估计值。因此，SNP 的效应和标准误的估计值可能是有偏的。不建议将上述软件所得结果用于 Meta 分析[31]。(2) 无论是 Firth 校正还是鞍点近似校正的统计量与经典统计量完全不同，包括：效应值、标准误、检验统计量和 P 值。因此，使用基于汇总数据(summary-level data)的统计方法时，需要事先确认可行性，包括但不限于：计算 SNP 遗传度、构建多基因评分(polygenetic risk score, PRS)等。(3) 本文中所列举的方法是均针对单位点(single-variant)的假设检验，获得的系数是该位点的边际效应，因此不宜直接将回归系数用于构建预测模型。唯一可用于构建预测模型的 R 软件包 *logistf* 也需要在 Firth 校正系数的基础上额外校正截距项[30]。

目前极端不平衡数据的 GWAS 统计分析已日臻完善。除了前文所提及用于二分类和生存数据的方法，鞍点近似方法还可拓展至多分类结局分析(POLMM[32])、基因-环境交互作

用分析(SPAGE[33])以及基因集(gene-based)分析(fastSPA-GENE[34]、POLMM-GENE[35]、SAIGE-GENE[36]、SAIGE-GENE+[37])。丰富多样的统计方法满足了各类 GWAS 数据分析的需求，实现了更合理的数据分析，为正确鉴定新的生物标志物提供了理论保障和技术支持。

参考文献

[1] ZHOU W, NIELSEN J B, FRITSCHE L G, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies [J]. Nat Genet, 2018, 50(9): 1335-41.

[2] DENNY J C, BASTARACHE L, RITCHIE M D, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data [J]. Nat Biotechnol, 2013, 31(12): 1102-10.

[3] DEY R, ZHOU W, KIISKINEN T, et al. Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks [J]. Nat Commun, 2022, 13(1): 5437.

[4] MA C, BLACKWELL T, BOEHNKE M, et al. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants [J]. Genet Epidemiol, 2013, 37(6): 539-50.

[5] DAI X, FU G, ZHAO S, et al. Statistical Learning Methods Applicable to Genome-Wide Association Studies on Unbalanced Case-Control Disease Data [J]. Genes (Basel), 2021, 12(5).

[6] LEHMANN E L, ROMANO J P, CASELLA G. Testing statistical hypotheses [M]. Springer, 1986.

[7] BERGER R L, CASELLA G. Statistical inference [M]. Duxbury, 2001.

[8] 韩栋, 陈征, 陈平雁, et al. 轮廓似然函数及其应用 [J]. 中国卫生统计, 2012, 29(04): 478-80+83.

[9] EFRON B, HINKLEY D V. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information [J]. Biometrika, 1978, 65(3): 457-83.

[10] KENDALL M, STUART A, ORD J, et al. Vol. 2A: Classical inference and the linear model [J]. London [etc]: Arnold [etc], 1999.

[11] DEY R, SCHMIDT E M, ABECASIS G R, et al. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS [J]. Am J Hum Genet, 2017, 101(1): 37-49.

[12] DEY R, NIELSEN J B, FRITSCHE L G, et al. Robust meta-analysis of biobank-based genome-wide association studies with unbalanced binary phenotypes [J]. Genet Epidemiol, 2019, 43(5): 462-76.

[13] COX D R. Partial likelihood [J]. Biometrika, 1975, 62(2): 269-76.

[14] MOORE D F. Applied survival analysis using R [M]. Springer, 2016.

[15] COX D R, SNELL E J. A general definition of residuals [J]. Journal of the Royal Statistical Society: Series B (Methodological), 1968, 30(2): 248-65.

[16] KANG H M, SUL J H, SERVICE S K, et al. Variance component model to account for sample structure in genome-wide association studies [J]. Nat Genet, 2010, 42(4): 348-54.

[17] YANG J, LEE S H, GODDARD M E, et al. GCTA: a tool for genome-wide complex trait analysis [J]. Am J Hum Genet, 2011, 88(1): 76-82.

[18] LOH P R, TUCKER G, BULIK-SULLIVAN B K, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts [J]. Nat Genet, 2015, 47(3): 284-90.

[19] BRESLOW N E, CLAYTON D G. Approximate inference in generalized linear mixed models [J]. Journal of the American statistical Association, 1993: 9-25.

[20] GILMOUR A R, THOMPSON R, CULLIS B R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models [J]. Biometrics, 1995: 1440-50.

[21] BI W, FRITSCHE L G, MUKHERJEE B, et al. A Fast and Accurate Method for Genome-Wide Time-to-Event Data Analysis and Its Application to UK Biobank [J]. Am J Hum Genet, 2020, 107(2): 222-33.

[22] FIRTH D. Bias reduction of maximum likelihood estimates [J]. Biometrika, 1993, 80(1): 27-38.

[23] GREENLAND S, MANSOURNIA M A. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions [J]. Statistics in medicine, 2015, 34(23): 3133-43.

[24] HEINZE G. The application of Firth's procedure to Cox and logistic regression [M]. Technical report 10. Department of Medical Computer Sciences, Section of Clinical Biometrics …. 1999.

[25] HEINZE G, SCHEMPER M. A solution to the problem of separation in logistic regression [J]. Stat Med, 2002, 21(16): 2409-19.

[26] LUGANNANI R, RICE S. Saddle point approximation for the distribution of the sum of independent random variables [J]. Advances in applied probability, 1980, 12(2): 475-90.

[27] FEUERVERGER A. On the empirical saddlepoint approximation [J]. Biometrika, 1989, 76(3): 457-64.

[28] HOSMER JR D W, LEMESHOW S, STURDIVANT R X. Applied logistic regression [M]. John Wiley & Sons, 2013.

[29] HOSMER JR D W, LEMESHOW S, MAY S. Applied survival analysis: regression modeling of time-to-event data [M]. John Wiley & Sons, 2008.

[30] PUHR R, HEINZE G, NOLD M, et al. Firth's logistic regression with rare events: accurate effect estimates and predictions? [J]. Stat Med, 2017, 36(14): 2302-17.

[31] MBATCHOU J, BARNARD L, BACKMAN J, et al. Computationally efficient whole-genome regression for quantitative and binary traits [J]. Nat Genet, 2021, 53(7): 1097-103.

[32] BI W, ZHOU W, DEY R, et al. Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes [J]. The American Journal of Human Genetics, 2021, 108(5): 825-39.

[33] BI W, ZHAO Z, DEY R, et al. A fast and accurate method for genome-wide scale phenome-wide G× E analysis and its application to UK biobank [J]. The American Journal of Human Genetics, 2019, 105(6): 1182-92.

[34] ZHAO Z, BI W, ZHOU W, et al. UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test [J]. The American Journal of Human Genetics, 2020, 106(1): 3-12.

[35] BI W, ZHOU W, ZHANG P, et al. Scalable mixed model methods for set-based association studies on large-scale categorical data analysis and its application to exome-sequencing data in UK Biobank [J]. Am J Hum Genet, 2023, 110(5): 762-73.

[36] ZHOU W, ZHAO Z, NIELSEN J B, et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts [J]. Nature genetics, 2020, 52(6): 634-9.

[37] ZHOU W, BI W, ZHAO Z, et al. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests [J]. Nat Genet, 2022, 54(10): 1466-9.